

Data: 12/05/2021

Assignatura: Visualització de dades



TREBALL DE SHINY

ESTADÍSTIQUES I PREDICCIONS DE LA NBA



GUILLEM MIRALLES



ÍNDEX DE CONTINGUTS:

1.	RESUM: OBJECTIUS DEL TREBALL	1
2.	ESTADÍSTIQUES GENERALS.....	2
	2.1. ESTADÍSTIQUES GENERALS DELS JUGADORS:	2
	2.2. ESTADÍSTIQUES GENERALS DELS EQUIPS:	3
3.	ESTADÍSTIQUES AVANÇADES	4
	3.1. ESTADÍSTIQUES AVANÇADES DELS EQUIPS:	4
	3.2. ESTADÍSTIQUES AVANÇADES DELS JUGADORS:	6
	3.3. ESTADÍSTIQUES AVANÇADES DELS LLANÇADORS:	7
4.	MODEL DE PREDICCIÓ:	8
5.	SALARI MITJÀ:	10
6.	DADES	11

Enllaç al projecte: https://guillemmiralles.shinyapps.io/5_42/

1. RESUM: OBJECTIUS DEL TREBALL

L'objectiu d'aquest treball es basa en realitzar una aplicació de Shiny que ens mostra dades sobre estadístiques y prediccions de la competició NBA. La NBA és una competició que genera moltíssimes estadístiques y es un dels esports on més ús es fa del BIG Data des de fa molts anys. Abans del projecte, es definiren varies idees per a dur a terme:

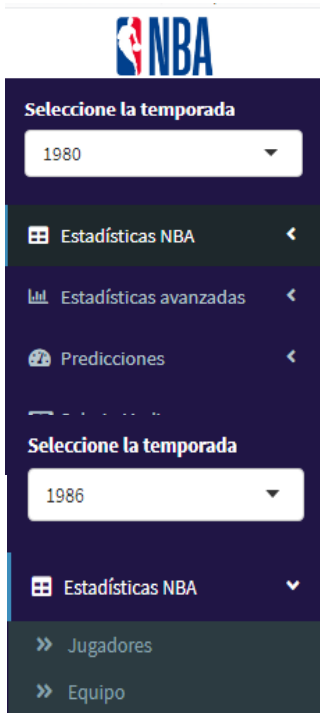
- **Estadístiques generals:** Informació general sobre equips i jugadors en una determinada temporada de l'NBA.
- **Estadístiques avançades:** Diversos gràfics, que aconseguixen explicar conceptes avançats sobre la NBA.
- **Models de predicció:** Diversos models de predicció que elaboren en una predicció sobre el que seran els jugadors del millor quintet d'un any determinat en l'NBA.
- **Salari Mitjà:** Com aquests esports cada vegada estan més mercantilitzats, serà interessant veure com evoluciona els salaris d'aquests jugadors.

Posteriorment, anirem definint els conceptes amb més detall. M'agradaria afegir que hi hauria moltíssima més informació que visualitzar, però el que s'ha intentat en aquest treball es crear una xicoteta mostra del que podria ser una aplicació més completa enfocada a complir tots els aspectes que es demanen en l'assignatura.

Per a l'elaboració de l'aplicació hem utilitzat **ShinyDashboard**, perquè investigant un poquet, vaig arribar a la [web d'exemples de ShinyDashboard](#), on vaig vore que més o menys era el que volia aconseguir, crec que es més similar al que seria una aplicació del smartphone i l'ordinador.



2. ESTADÍSTIQUES GENERALS



Per començar, vam introduir els diferents menús que ja estaven pensat com he comentat abans. Primerament, amb les dades de [Season Stats](#), vem crear el **SelectInput** de la temporada, aquest el vam posar al servidor per a que les entrades que es pogueren elegir estigueren dins del dataframe de estadístiques de jugadors. Ho fem de la següent forma:

```
output$id_season <- renderUI({selectInput('season', 'seleccione la temporada',  
choices = unique(df_ad %>% dplyr::select(yearSeason))
```

Aleshores, el `id_season` es el que posem al ui amb la instrucció **uiOutput('id_season')**, mentre que **input\$season** serà el que utilitzem al server per a fer referència a l'entrada de l'usuari.

Dins del **sidebarMenu**, on prèviament hem posat el **selectInput** i el menulitem de Estadístiques de la NBA, ara afegim els tres **menuSubItem**, Jugadors i Equip. Aquests dos submenús si que tindran contingut al **dashboardBody**, on posarem tables dinàmiques que contindran la informació desitjada.

2.1. ESTADÍSTIQUES GENERALS DELS JUGADORS:

```
output$tabla_temporada <- DT::renderDataTable({  
  df_p %>% subset(yearSeason==input$season) %>%  
  dplyr::select(namePlayer, slugPosition, agePlayer, slugTeamBREF,  
    minutesTotals, trbTotals, astTotals, ptsTotals,  
    blkTotals, stlTotals)  
})
```

Ara, amb el dataframe `df_p` que son les dades de `Season_Stats`, el que fem es un subset de l'entrada elegida per l'usuari, i posteriorment de les moltes estadístiques que té el dataframe, exactament 53, elegim les més rellevants per als jugadors ja que estem fent estadístiques generals.

Amb quines ens quedem?

- **NamePlayer** : Nom del jugador.
- **SlugPosition**: Posició.
- **agePlayer**: Edat.
- **slugTeamBREF**: Equip.
- **minutesTotals**: Minuts totals en la temporada seleccionada.
- **trbTotals**: Rebots Totals en la temporada seleccionada.
- **astTotals**: Assistències totals en la temporada seleccionada.
- **ptsTotals**: Punts totals en la temporada seleccionada.
- **blkTotals**: Bloquejos totals en la temporada seleccionada.
- **stlTotals**. Robatoris totals en la temporada seleccionada.



Simplement, per afegir-ho al ui, anem al dashboardBody, fem referència a 'player', que es el submenú dels jugadors, i afegim la taula dinàmica amb DT:DTOutput('tabla_temporada'):

```
dashboardBody(tabItems(
  tabItem('player', h2("Estadísticas de los jugadores de la NBA en la temporada seleccionada:"),
    DT::DTOutput('tabla_temporada'),
  ),
```

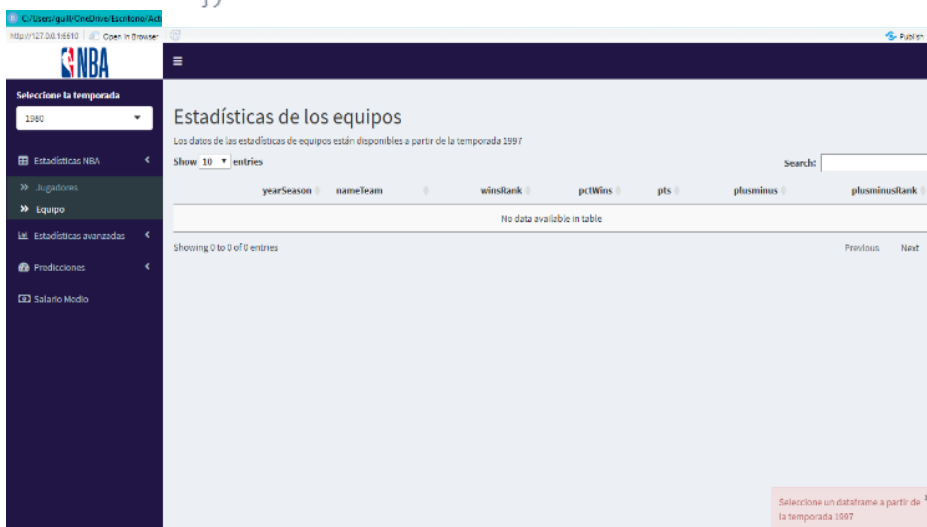
2.2. ESTADÍSTIQUES GENERALS DELS EQUIPS:

Aquest procés es prou semblant al anterior, donat el [dataframe de estadístiques dels equips](#).

```
df_t %>% subset(yearSeason==input$season) %>%
  dplyr::select(yearSeason, nameTeam, winsRank, pctWins, pts, plusminus, plusminusRank)
})
```

Com podem vore, fem el mateix a l'anterior, definim la funció DT::renderDataTable, i dins el que fem es que **donat el any que haja seleccionat el usuari extraiem un dataset i elegim les estadístiques que volem mostrar**. El problema es que ací l'usuari **no pot elegir qualsevol any**, perquè les estadístiques generals de equips **sols estan a partir de l'any 1997**, llavors la idea ara és crear **un sistema que llanci una notificació a l'usuari si elegeix un any que no esta disponible**.

```
output$tabla_temporada_eq <- DT::renderDataTable({
  observeEvent(input$season, {
    if (input$season < 1997){
      id <- showNotification(paste("Seleccione un dataframe a partir de la temporada 1997"),
        type = 'error', closeButton=TRUE)
      ids <- c(ids, id)}
    })
  observeEvent(input$season, {
    if (input$season > 1996){
      while (length(ids) > 0){
        removeNotification(ids[1])
        ids <- ids[-1]}}
    })
})
```



Com podem veure, establim un error que es llança si l'usuari elegeix un any inferior al 1997, i creem la variable id que es el error. Aquest error el tenim que eliminar. Aleshores el que fem més avall es que si selecciona un dataframe correcte, elimina totes les notificacions que ses guarden a la variable ids.

Data: 12/05/2021

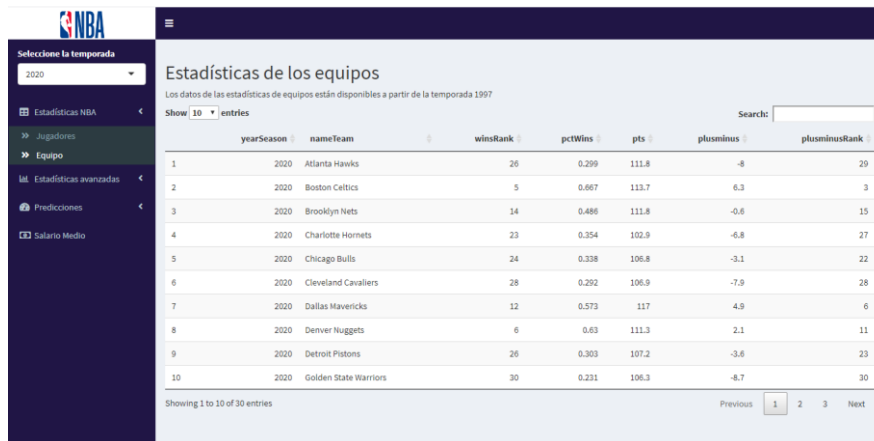
Assignatura: Visualització de dades

Autor: Guillem Miralles



Com podem veure, llança l'excepció que hem dit, i a demés la tabla ens indica que no hi ha dades disponibles. Per a influir tot açò al ui, fem el mateix que amb les estadístiques generals dels jugadors.

La taula final seria la següent:



	yearSeason	nameTeam	winsRank	pctWins	pts	plusminus	plusminusRank
1	2020	Atlanta Hawks	26	0.299	111.8	-8	29
2	2020	Boston Celtics	5	0.667	113.7	6.3	3
3	2020	Brooklyn Nets	14	0.486	111.8	-0.6	15
4	2020	Charlotte Hornets	23	0.354	102.9	-6.8	27
5	2020	Chicago Bulls	24	0.338	106.8	-3.1	22
6	2020	Cleveland Cavaliers	28	0.292	106.9	-7.9	28
7	2020	Dallas Mavericks	12	0.573	117	4.9	6
8	2020	Denver Nuggets	6	0.63	111.3	2.1	11
9	2020	Detroit Pistons	26	0.303	107.2	-3.6	23
10	2020	Golden State Warriors	30	0.231	106.3	-6.7	30

Aquesta tabla conté informació com els punts de l'equip en eixa temporada o el percentatge de victòries.

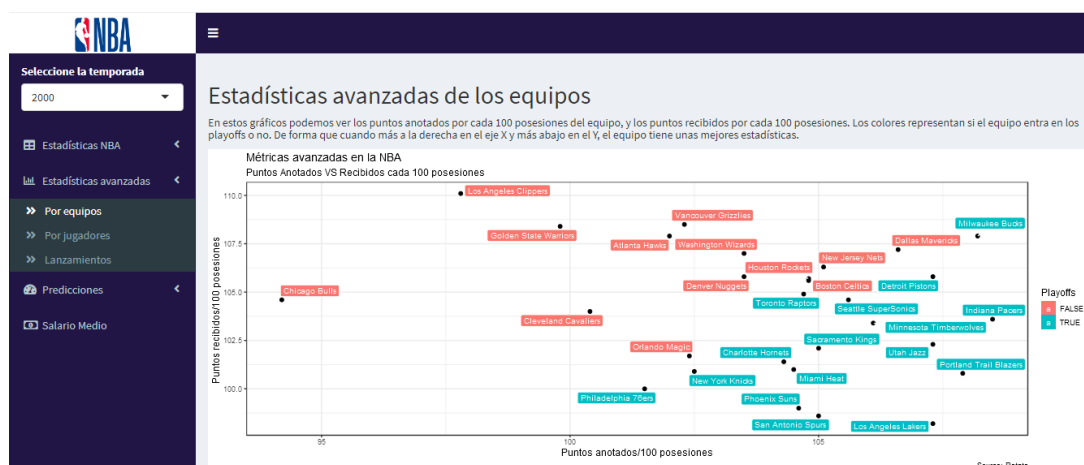
Igual que amb la dels jugadors, també es pot buscar un equip al buscador de la taula.

3. ESTADÍSTIQUES AVANÇADES

Com hem dit abans, aquesta secció està enfocada a visualitzar informació avançada de la NBA. Intentem explicar amb gràfics conceptes que, a simple vista (mirant una taula) són imperceptibles.

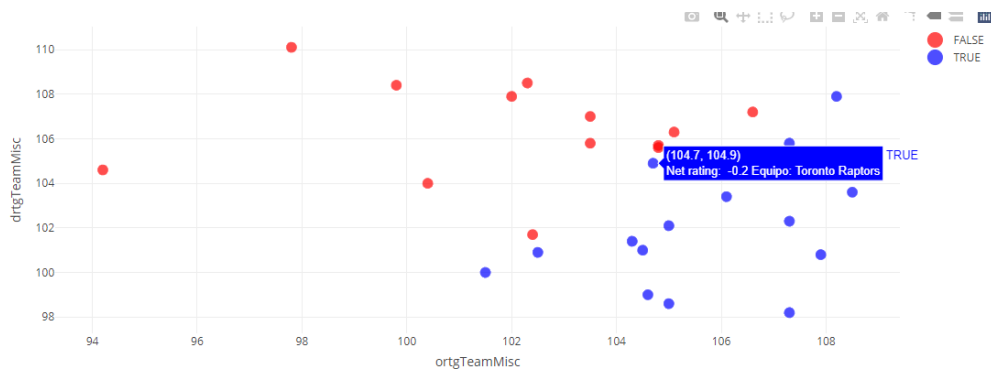
3.1. ESTADÍSTIQUES AVANÇADES DELS EQUIPS:

Donada una temporada seleccionada a l'input inicial, i les dades d'[estadístiques avançades d'equips](#), elaborem dos gràfics, un en ggplot i l'altre amb plotly que representen el mateix. En aquests gràfics podem veure els punts anotats per cada 100 possessions de l'equip, i els punts rebuts per cada 100 possessions. Els colors representen si l'equip entra en els playoffs o no. De manera que quan més a la dreta en l'eix X i més avall en el Y, l'equip té unes millors estadístiques.





En la majoria d'anys veiem una línia de tall clara que divideix als equips que entren en playoffs i als que no. Açò pot ajudar a un equip a estimar molts punts ha de rebre per cada punt a favor per a entrar al playoffs.



El segon gràfic que apareix a aquest submenú, es interactiu amb plotly, on podem ampliar per a avaluar zones que ens puguen interessar, o vore els valors exactes de l'equip.

```
output$ggplot1 <- renderPlot({
  df_ad <- df_ad %>% subset(yearSeason==input$season)

  ggplot(data = df_ad, aes(ortgTeamMisc, drtgTeamMisc)) +
    geom_point(color = 'black') +
    theme_bw() +
    geom_label_repel(aes(label = nameTeam,
                        fill = factor(isPlayoffTeam)),
                    color = 'white',
                    size = 3.5) + xlab("Puntos anotados cada 100 posesiones") + ylab ("Puntos recibidos cada 100 posesiones")
  labs(title = "Métricas avanzadas en la NBA",
        subtitle = "Puntos Anotados VS Recibidos cada 100 posesiones",
        caption = "Source: Rstats",
        x = "Puntos anotados/100 posesiones",
        y = "Puntos recibidos/100 posesiones",
        fill = "playoffs"))

output$plotly2 <- renderPlotly({
  df_ad <- df_ad %>% subset(yearSeason==input$season)
  pal <- c('red','blue')
  plot_ly(data = df_ad, x = ~ortgTeamMisc, y = ~drtgTeamMisc, color = ~isPlayoffTeam,
          size = 16, colors = pal,
          text = ~paste("Net rating: ", nrtgTeamMisc, 'Equipo:', nameTeam))
})
```

Primer creem un **subset** del **dataframe** que volem representar donada la entrada elegida per l'usuari.

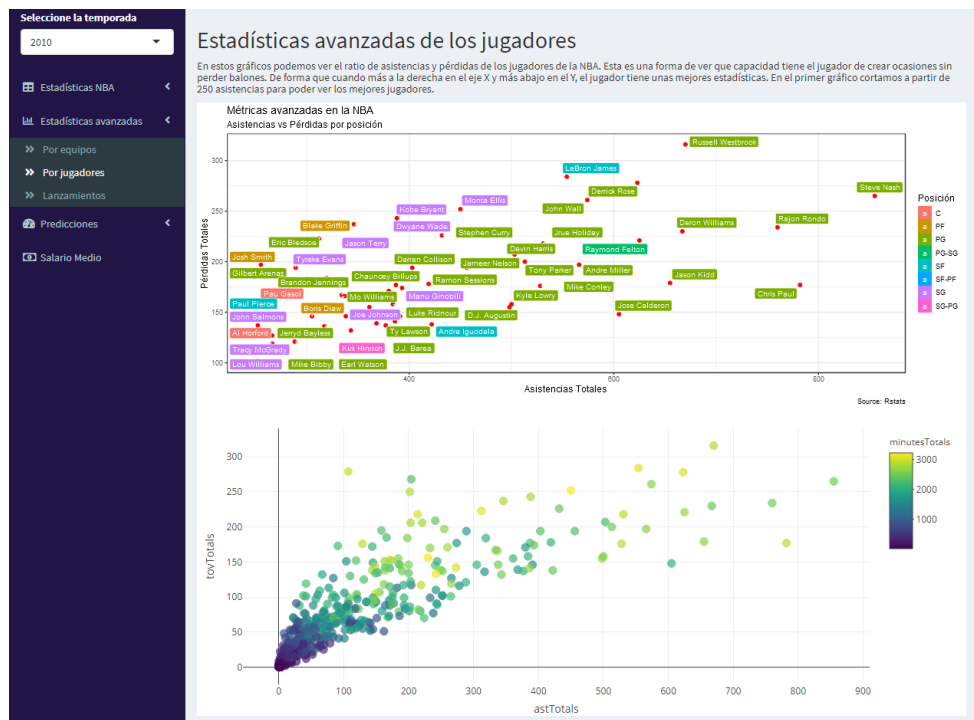
Posteriorment **elaborem el ggplot**, al aes del plot fiquem les estadístiques a representar, i **geom_label_repel** ens ajuda a que al plot apareguen els noms dels equips.

En el **output\$plotly2**, simplement creem el gràfic interactiu que volem representar.

En el ui fem un **plotOutput(ggplot1)** i un **plotlyOutput(plotly2)**. Afegim un **%>%withSpinner()** ja que, una volta seleccionat aquest menú, ens mostra una icona de com carrega la pàgina.



3.2. ESTADÍSTIQUES AVANÇADES DELS JUGADORS:

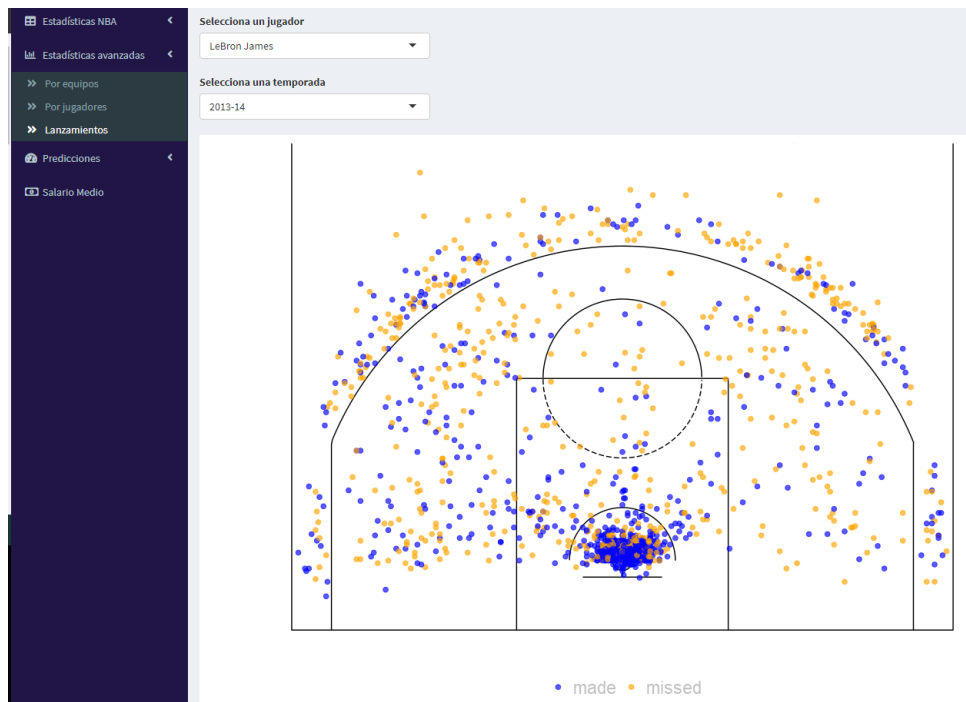


En aquests gràfics podem veure el ràtio d'assistències i pèrdues dels jugadors de la NBA en una temporada determinada. Aquesta és una forma de veure que **capacitat té el jugador de crear ocasions sense perdre balons**. De manera que quan més a la dreta en l'eix X i més avall en el Y, el jugador té unes millors estadístiques. En el primer gràfic tallem a partir de 250 assistències per poder veure els millors jugadors. Els colors representen la posició dels jugadors.

Com es podrà comprovar, si canviem la temporada, la majoria de jugadors que hi ha a la són de color verd, que representa la posició de base, açò és comprensible ja que els base són els encarregats de distribuir el joc. Aleshores es veu que bases han fet una millor temporada i han creat un major nombre d'ocasions. La metodologia per a crear els dos gràfics i afegir-los a la ui es prou paregut a l'anterior.



3.3. ESTADÍSTIQUES AVANÇADES DELS LLANÇADORS:



Donat un dels jugadors disponibles a les [dades](#), els gràfic mostra els tirs encestats i fallats per aquest jugador, en una temporada seleccionada, i la posició des d'on l'ha realitzat. Aquesta gràfica pot ser útil per a veure que estil de joc té cada jugador i des d'on es més o menys efectiu. Per exemple, si escollim a Curry, veurem com es un gran llançador de triples, mentre si seleccionem a LeBron, veurem com la gran part son de prop de la cistella.

```
output$id_player <- renderUI({selectInput('shot_player','Selecciona un jugador',
  choices = unique(nba_shots %>% dplyr::select(player_name))}))

output$id_season_shots <- renderUI({selectInput('shot_season','Selecciona una temporada',
  choices = unique(nba_shots %>% dplyr::select(season))}))

output$plot_shots <- renderPlot({
  source("helpers.R")
  gg_court = make_court()
  player_data = filter(nba_shots, player_name == input$shot_player, season == input$shot_season)
  gg_court + geom_point(data = player_data, alpha = 0.65, size = 2.3,
    aes(loc_x, loc_y, color = shot_made_flag)) +
    scale_color_manual("", values = c(made = "blue", missed = "orange"))
})
```

Primerament, donades les dades de [Nba Shots](#), creem dos select inpputs, uno per als jugadors disponibles i altre per a les temporades. Aquest gràfic es possible gràcies a un fitxer, 'helpers.R', que es pot trobar a l'enllaç:

http://juliawrobel.com/tutorials/shiny_tutorial_nba.html, i es el encarregat de crear la figura de la pista de bàsquet (**gg_court**), on s'afegeixen els diferents punts. Aquesta part esta explicada a tutorial de l'enllaç. Després, afegim el plot_shots al ui i ja estaria aquesta part creada.



4. MODEL DE PREDICCIÓ:

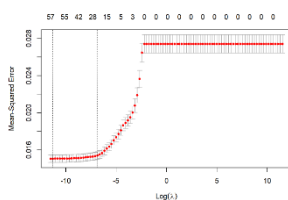
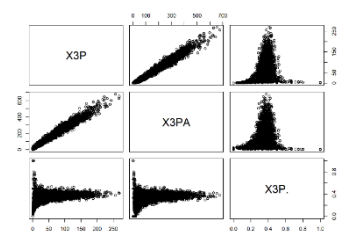
Els models de predicció intenten predir el millor quintet de la NBA d'una determinada temporada en base a les estadístiques dels jugadors. Que és el millor quintet de la NBA?

El **millor quintet de la NBA és un premi anual de la NBA atorgat als millors jugadors** de la temporada. Un grup de periodistes i locutors esportius dels Estats Units i del Canadà voten aquest premi, el qual consta de tres equips de cinc jugadors (primer equip, segon equip i tercer equip). Voten 5 jugadors per al primer equip, 5 per al segon i 5 per al tercer. **Els cinc jugadors amb un nombre total més gran de punts entren al primer equip, el segon pels següents cinc jugadors i el mateix amb el tercer equip. Però hi ha una restricció: la posició.** En cada vot de 5 jugadors (que componen un equip), 2 jugadors són de escolta (al nostre quadre de dades "PG" i "SG"), els altres 2 són alers ("SF" i "PF") i l'últim jugador és el base ("C"). Aquests són bàsicament els 15 millors jugadors de la temporada.

El model s'elabora amb les dades de **Season_Stats.csv**, que conté les estadístiques dels jugadors en cada temporada. Algunes són simples com els tirs de 3 o els minuts jugats, i altres són més avançades, com l'aportació d'un jugador defensivament per cada 100 possessions. Com sabem, algunes seran més rellevants per al nostre model que altres.

Els passos seguits per a l'elaboració del model son els següents:

- **Lectura i neteja de les dades** (lectura del CSV, configuració del tipus de variable, canvi de valors nuls a 0, eliminació de files amb reproductors duplicats ...).
- **Introduir una nova variable** en les dades que ens indique si el jugador ha estat al millor quintet de la temporada o no.
- **Creació d'un conjunt train i un conjunt de test.** El conjunt de formació tindrà dades del 1980 al 2011 i el conjunt de proves tindrà dades del 2011 al 2017 (80% - 20%).
- Visualitzem les dades i observem que **moltes de les variables es correlacionen entre si** o no ens proporcionen informació rellevant. Com en aquest cas, el total de triples encestats, el total de triples encestats, i el percentatge encestats/tirats.
- Per saber quines utilitzarem, utilitzarem tècniques de reducció que ens ajuden a trobar les millors variables per al nostre model. **La reducció de dimensionalitat amb LASSO** ens ajuda a seleccionar les millors variables.

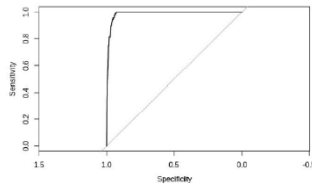


(Intercept)	PosPF	PosSF	Age	G	MP	TS.
2.222875e-02	-1.436547e-03	-3.706560e-03	3.249107e-04	7.727267e-04	-1.716349e-04	-4.249438e-02
FTr	STL.	TOV.	USG.	DWS	WS	BPM
-1.607393e-02	-7.022702e-04	2.898649e-04	-2.537874e-04	6.191524e-03	1.821565e-02	-4.986440e-04
VORP	FG	FGA	X2P	FTA	PF	ORB
3.161288e-02	2.369019e-05	1.295914e-04	6.016663e-05	3.683313e-04	-3.212804e-04	-1.507431e-04
DRB	AST	STL	BLK			
1.636141e-04	1.907179e-04	-9.944877e-05	3.400609e-04			

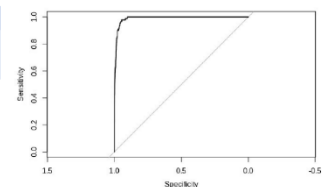


Veiem com les variables que ens interessin es redueixen molt (de 53 que hi ha, a 24). Com que realitzem una regressió logística, en les variables que hem obtingut del punt anterior, **realitzem tres models utilitzant tres mètodes diferents que serem els que compararem.** Aquests tres mètodes són: Regressió Logística Múltiple (**GLM**), Anàlisi Discriminant Quadràtic (**QDA**) i Anàlisi Discriminant Lineal (**LDA**). No prenem el mètode KNN perquè ja sabem que els valors veïns no són interessants per predir el valor següent.

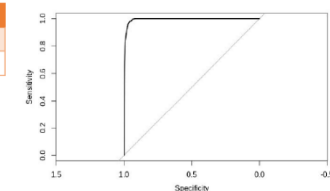
QDA	0	1
0	3296	4
1	173	86
% Hits	95.02669%	
Area under the curve	0.9881	
IC 95%	0.9841-0.992	



LDA	0	1
0	3422	21
1	47	69
% Hits	98.08935%	
Area under the curve	0.9901	
IC 95%	0.9862-0.994	



GLM	0	1
0	3462	37
1	7	53
% Hits	98.7637 %	
Area under the curve	0.9946	
IC 95%	0.9934-0.9958	

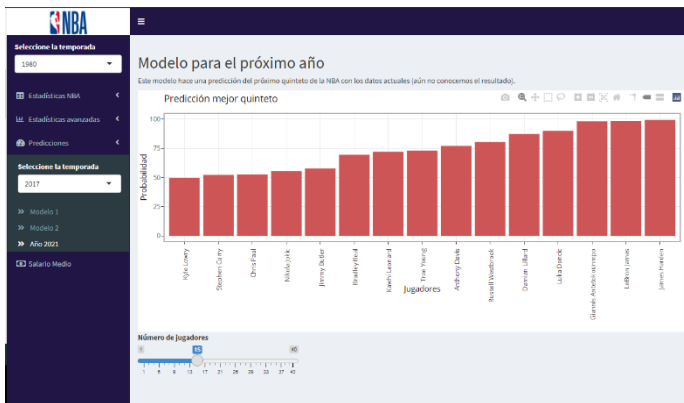
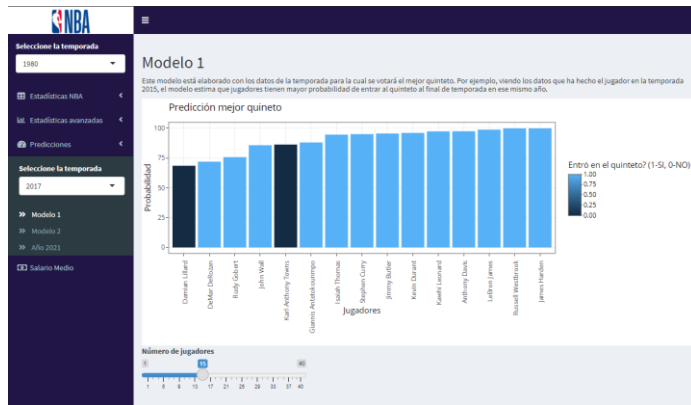


Triem el **mètode GLM**, ja que és el que **millor prediu els vertaders positius i negatius**. El que volem és el **mínim error de falsos positius**, cosa que ens interessa per al model, però els falsos negatius són superiors als altres models.

- **Amb les prediccions del model, creem un dataframe** que conté: la probabilitat que els jugadors formen part del quintet, si realment està en el quintet, el nom, la posició. **Però tenim un problema en aquests equips hi ha una restricció de posició.** Per aquest motiu, **crearem una funció que selecciona els jugadors per formar un equip** (2 escoltes, 2 aleros i 1 pivot). La funció que fa és triar (en primer lloc) 2 escoltes amb més probabilita, 2 aleros i 1 base (per al primer equip), i repeteix el procés amb el segon equip i el tercer.

Aleshores, ja tindríem el model. Però, per què hi ha dos models de predicció?

- **Model 1:** Aquest model està **elaborat amb els dades del final de temporada per a la qual es votarà el millor quintet**. Per exemple, veient els dades que ha fet el jugador en la temporada 2015, el model estima que jugadors tenen major probabilitat d'entrar al quintet a la final de 2015. Té una major precisió que el model 2 a que ho estima amb un menor interval de temps.
- **Model 2:** Aquest model està **elaborat amb els dades de la temporada anterior a la votació del millor quintet**. Per exemple, veient els dades que ha fet el jugador en la temporada 2014, el model estima que jugadors tenen major probabilitat d'entrar al quintet del any següent, 2015. Té un interval de temps major que el de model 1.
- **Model 2021:** Aquest model utilitza les **dades actuals de la temporada de la NBA** per a estimar el quintet d'aquest any.



Com podem veure, tenim un **selectInput** que està creat al ui per als anys 2015, 2016 i 2017. Aquests anys són els que el model elabora una predicció. També tenim un **sliderInput**, que per defecte està en 15 (jugadors del quintet), però es pot **allargar per a veure altres jugadors als que el model dona una alta probabilitat**. Aquests gràfics estan elaborats amb plotly i es pot interactuar per a veure la probabilitat del jugador.

Amb color blau clar podem veure els que el model ha encertat i amb blau fosc els que no. També podem dir que el model 1 funciona prou bé ja que, de 300 jugadors que disputen una temporada de la NBA, sempre encerta més de 10.

5. SALARI MITJÀ:

Aquest gràfic és prou bàsic. Amb les dades de salaris, elaborem amb plotly un **gràfic interactiu** de la evolució de la mitjà de diners destinats pels equips de la NBA al salari dels jugadors.





6. DADES

Dades de Est. Generals De Jugadors i utilitzades per a la elaboració dels models:

<https://www.kaggle.com/drgilermo/nba-players-stats> (fitxer Season_Stats.csv)

Dades de Est. Generals De Equips:

Utilitzem la llibreria **nbastatR**: <http://asbcllc.com/nbastatR/>

```
nbastatR::teams_annual_stats
```

Amb la funció anterior, R carrega les dades en el medi directe. Però les hem guardat creant un fitxer perquè era més fàcil. El fitxer es: [teams.csv](#)

Dades de Est. Avançades (per jugadors y equips):

Utilitzem la llibreria **nbastatR**: <http://asbcllc.com/nbastatR/>

Per a jugadors: `nbastatR::players_tables` (fitxer [playerstotals.csv](#))

Per a equips: `nbastatR::teams_tables` (fitxer [teamsadvanced.csv](#))

ADEstadístiques avançades - Llançadors:

http://juliawrobel.com/tutorials/shiny_tutorial_nba.html

Des d'aquesta pàgina, ens descarreguem un zip d'on tenim:

- **nba_shots.RData**: Informació dels tirs que utilitzem en aquesta secció
- **helpers.R** : Figura del camp de bàsquet que utilitzem per a fer el plot

Utilitzem: `load("nba_shots.RData")`

Dades del models:

Per a entrenar el model y realitzar la predicció utilitzem les dades de Season_Stats (amb la variable quintet introduïda).

Extracció amb `write_csv` guardem les eixides de cada model.

Fitxers: `df_mod1` (model 1), `df_mod2` (model 2), `df_mod3` (any 2021)

Dades de salaris:

<https://data.world/datadavis/nba-salaries> (fitxer [dbsal.csv](#))